University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Omar G. Soto Fortuño TOTAL: /100

Student ID: 843-07-8140

Section: ____________ June 29, 2017

Answer the following questions.

I) **(25 pts)** **Problem I: MapReduce**: Briefly explain the concept of MapReduce, and provide an example on what problems you can solve with it.

MapReduce is a model to work with Big Data. It consists of two phases: Map and Reduce. There can be an intermediate phase that is called Combine.

They idea about this is that the Mapper, will receive a chunk of a data-set to be analyzed. The Mapper will process that and extract a key-value pair. After all the Mappers are done with their process, the Reducers will come into action. In MapReduce there is a warranty that a single Reducer will receive all the data about a single key, that will allow the Reducer to finally process (mostly summarize) that data and give a result per key.

An example of one problem that can be solved with this is to count screeName appearrances on a tweets data-set…

The input will be a bunch of tweets in JSON. The mapper will just return a key-value pair with ([Screen Name], 1) for each occurrence of that screenName within his data. The reducers will receive then a bunch of data associated with a particular key, for example:
(Omar, 1)
(Omar, 1)
(Omar, 1)

And… Using that, will output (Omar, 3). This will happen for all the available keys. In this example I just displayed one key but It can't be a lot of keys and a particular Reducer can process several keys.

II) **(25 pts)Problem II: Spark RDD**: Briefly explain the concept of the Spark RDDs and contrast their performance with that of MapReduce.

A Spark RDD (Resilient Distributed Dataset) is an object that can be created using deterministic operations based on storage data (for example, Hadoop) or based on a previously defined RDD. It's great because you can chain several RDDs and only the first one point directly to the data. The RDD is computed when is needed, when that happens it optimize which internal operations it needs to give the output (depending on the RDD or the chain of RRDs). It implements MapReduce but internally, after optimizing what to do in MR.

It's better to use this instead of MR because, you will need to code a bunch of Mappers and Reducers and do several MR processes to get an output while Spark RRD while optimize your request and then execute a MR process or a series of MR processes, depending on the situation.

III) **(25 pts)** **Problem III: Supervised Learning**: Briefly explain the concept of Supervised Learning, and provide an example on what problems you can solve with it.

Supervised Learning is a Machine Learning area that creates a model based on a training data-set that contains an input and the desired output. This is often use on classification problems or in regression problems.

A classification problem example will be to classify tweets between Positive and Neutral. A training data-set is required that contains the tweets and it respective prediction. Using that, the ML will implement a supervised learning algorithm and create a model for it. The desired output is: (Tweet, [Positive or Negative])

A regression problem example will be to get the coffee sales depending on the timeframe. A training data-set is required that contains previous days sales of coffee per timeframe. Using that, the ML will implement a supervised learning algorithm and create a model for it. The desired output is: (Timeframe, [Quantity of Coffee Sales])

IV) **(25 pts)**<u>**Problem IV: Unsupervised Learning**</u>: Briefly explain the concept of Unsupervised Learning, and provide an example on what problems you can solve with it.

Unsupervised Learning is a Machine Learning area that creates a model by having inferences about input data that doesn't have a desired output (or prediction). There are different Unsupervised Learning algorithms, on of them is based un clustering.
This is self-guided.

UL can be used to, based on a input of genomes, find clusters of that genomes that for example, can cause a common deseases. Or, to using a data-set of images, determine clusters of different areas:
water, terrain, roadways, urbanizations, etc.